Poster III-34

Biomarker and Molecular Diagnostic Discovery Using Genetic Programming

Worzel, W.P.*[*1], MacLean, C.D.[1], Driscoll, J.A.[2]
[1]Genetics Squared, Inc., Milan, MI, USA; [2]Middle Tennessee State University, Murfreesboro, TN, USA

It is well known that the most highly expressed genes are not always the most important genes in a disease process. However, most discovery methods based on gene expression profiles start by looking at the genes with the greatest variance. A novel machine learning technique is described that automatically selects genes and develops comprehensible rules relating genes to one another. In many cases the genes selected are not among the most highly expressed or the most highly variant. These genes have the potential to be significant new biomarkers for diseases or disease states. An example is given using small round blue cell tumor (SRBCT) data.

With the advent of cDNA microarrays and oligonucleotide arrays (together called 'gene chips' hereafter), it has become possible to observe variations in expression levels between healthy and diseased tissue or between different types of diseases. The National Cancer Institute and several research institutes have worked extensively in this area studying different cancers and their expression profiles. A typical study compares closely related cancers with a goal of discovering key differences for diagnostic and prognostic purposes and to gain insight into the regulatory and metabolic pathways associated with the targeted cancers. However, because of the large number of genes profiled and comparatively small number of samples available for such studies, it is often difficult to find the key differences between diseases.

Many analytic computing methods have been used in pursuit of this information including statistical methods such as regression analysis, visualization techniques such as cluster analysis and machine learning approaches such as neural networks, support vector machines (SVMs) and genetic algorithms. While there has been some success with these techniques, each has their own shortcomings. Most statistical methods have difficulty with non-linear processes, clustering techniques usually consider a relatively small number of genes based on fold variation and existing machine learning techniques usually deliver accurate, but "black box" solutions do not give much insight into the relationship between genes and their relative importance.

Genetic programming is a machine learning system that can automatically select a small number of genes from a gene chip having thousands or even tens of thousands of genes and combine them in human comprehensible computer programs in the form of mathematical expressions. The genes selected, their grouping and the relative importance of these genes in classifying diseases give an intriguing insight into the disease process and identify potential biomarkers for the targeted diseases. This poster describes the application of genetic programming to the problem of differentiating between Small, Round Blue-Cell Tumors (SRBCTs), assesses the generality of the rules within the data set and discusses possible biomarkers for each tumor-type studied.